ГЛАВА XY. Показатели качества в задачах чёткой бинарной классификации

Основной вопрос не какой алгоритм выбрать, и не как его обучить, а как оценить качество его работы.
В. Вапник

Единственная настоящая ошибка – не исправлять своих прошлых ошибок
Конфуций

В этой главе рассматриваем задачу классификации на два класса (с метками 0 и 1), классификатор выдаёт предполагаемую метку класса (т.е. «чёткий ответ»: какому классу принадлежит объект). На рис. XY.1 показано графическое представление размеченной выборки и решения[1] (полученного классификатора). Как и ранее, считаем, что у нас есть размеченная выборка $(x_1, y_1),(x_2, y_2),\ldots(x_m, y_m)$ (произвольная: это может быть обучающая выборка или отложенная тестовая) и ответы классификатора на объектах выборки: $a_1, a_2, \ldots, a_m$. Естественным, простым и распространённым показателем качества является **доля верных ответов или точность**[2] **(Accuracy, Accuracy Score** или **Mean Consequential Error):**

$$\text{Accuracy} = \frac{1}{m}\sum_{i=1}^{m} I[a_i = y_i],$$

т.е. просто доля объектов выборки, на которых алгоритм выдал правильные ответы. Недостаток такого функционала очевиден: он плох в случае дисбаланса классов (когда представителей одного из класса существенно больше, чем другого). В этом случае константный классификатор, который выдаёт метку большего класса формально имеет высокую точность. Это может не согласовываться с логикой решения задачи. Например, в задаче обнаружения

[1] Геометрически пространство делится на два подпространства (объекты из первого классификатор относит к классу 1, из второго – к классу 0). Здесь разделение производится прямой на плоскости.

[2] Термин точность также используется для перевода показателя качества Precision, поэтому возникает путаница. В последнее время часто «Accuracy» переводят как «аккуратность».

спама спамовых сообщений обычно сильно меньше, чем нормальных. Но алгоритм, который каждое письмо классифицирует как нормальное, т.е. не обнаруживает спам, нам как раз не нужен.

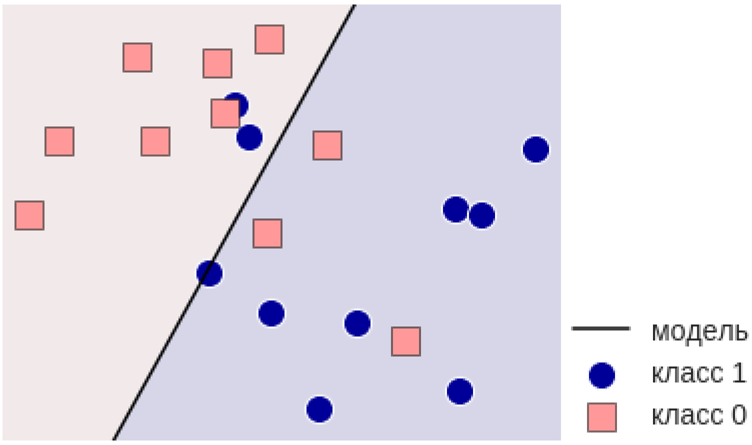


Рис. XY.1. Иллюстрация задачи с двумя классами и возможного решения.

Рассмотрим т.н. **матрицу несоответствий / ошибок (Confusion Matrix)** – матрицу размера[1] 2×2, ij-й элемент которой равен числу объектов i-го класса, которым алгоритм присвоил метку j-го класса.

В матрице несоответствий лежит полная информация об ошибках чёткого бинарного классификатора.

	$a=0$ negative	$a=1$ positive
$y=0$	7 True Negative	3 False Positive
$y=1$	2 False Negative	8 True Positive

Рис. XY.2. Матрица несоответствий.

На рис. XY.2 показана такая матрица для решения с рис. XY.1, также показаны названия элементов матрицы. В задаче бинарной классификации один из классов называют **положительным или позитивным** (обычно ему соответствует метка 1), другой – **отрицательным или негативным** (метка 0 или –1). Объекты, которые алгоритм относит к положительному классу, называются **положительными[2] (Positive)**, те из них, которые на самом деле

[1] Для случая 2х классов. Иногда матрицу несоответствий изображают по-другому: в транспонированном виде (ответы алгоритма соответствуют строкам, а правильные метки – столбцам). Здесь выбрана ориентация, которая соответствует реализации в библиотеке sklearn.

[2] Это довольно неестественно, логичнее было бы назвать объекты положительного класса 1 «положительными» (тем более, что «положительность» свойство объектов, а не алгоритма), но так принято и это согласуется с другими терминами, например TP – «истинно положительные», см. дальше.

принадлежат к этому классу – **истинно положительными (True Positive)**, остальные – **ложно положительными (False Positive)**. Аналогичная терминология есть для **отрицательного (Negative)** класса. Дальше используем естественные сокращения:

- TP = True Positive,
- TN = True Negative,
- FP = False Positive,
- FN = False Negative.

Для точности (Accuracy) справедлива формула:

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FN} + \text{TP} + \text{FP}}$$

	$a = 0$	$a = 1$
$y = 0$	TN	FP
$y = 1$	FN	TP

Здесь и далее для простоты запоминания формул показываем изображения матрицы несоответствий: тёмная закраска соответствует числителю, более светлая – знаменателю.

Ошибки классификатора делятся на две группы: первого и второго рода. В идеале (когда точность равна 100%) матрица несоответствий диагональная, ошибки вызывают отличие от нуля двух недиагональных элементов:

ошибка 1 рода (Type I Error) случается, когда объект ошибочно относится к положительному классу, равна FP/m,

ошибка 2 рода (Type II Error) случается, когда объект ошибочно относится к отрицательному классу, равна FN/m.

На рис. XY.3 показаны известные шуточные иллюстрации ошибок 1 и 2 рода.

Пример для простоты запоминания: студент сдаёт зачёт. Если он учил и знает, то принадлежит классу с меткой 1, иначе – имеет метку 0 (вполне логично называть знающего студента «положительным»). Экзаменатор выполняет роль классификатора: ставит зачёт (т.е. метку 1) или отправляет на пересдачу (метку 0). Самое желаемое для студента «не учил, но сдал» соответствует ошибке 1 рода, вторая возможная ошибка «учил, но не сдал» – 2 рода.

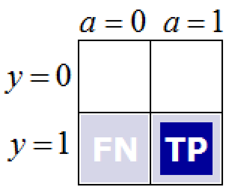
> Если студент не учил, но сдал (самое желаемое для студента), то экзаменатор допустил ошибку 1 рода, а наоборот (учил, но не сдал) – 2 рода.

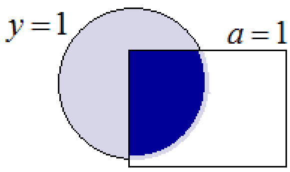
Рис. XY.3. Ошибка 1 рода (слева) и ошибка 2 рода (справа).

Через введённые выше обозначения выражаются следующие функции:

Полнота или чувствительность (Sensitivity, True Positive Rate, Recall, Hit Rate) – доля объектов положительного класса, которая верно классифицирована:

$$\mathrm{TPR} = R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$





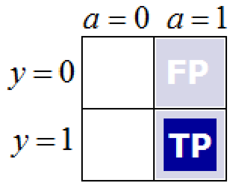
Полнота часто используется в информационном поиске. Когда пользователь вводит запрос, то существует множество релевантных ответов (ответы с меткой $y=1$), а на запрос алгоритм выводит перечень ответов, которые он посчитал релевантными ($a=1$). **Выдача называется полной**, если содержит все релевантные ответы (здесь как раз полнота равна 1). Если в выдаче присутствует только половина из множества релевантных ответов, тогда полнота равна 0.5. На рис. выше круг соответствует релевантным ответам, прямоугольник – выдаче[1].

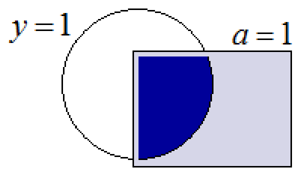
Точность (Precision, Positive Predictive Value) отражает, какой процент положительных объектов (т.е. тех, что мы считаем положительными) правильно классифицирован:

Отличайте точность «Accuracy» и точность «Precision». К сожалению, по-русски их часто называют одинаково.

[1] Надеемся, с интерпретацией рис. не будет проблемы (здесь пересечение сложно закрасить одновременно светлым и тёмным).

$$\mathrm{PPV} = P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$





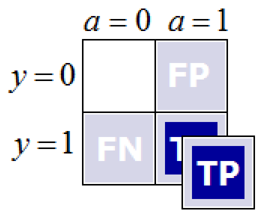
Точность также используется в информационном поиске и показывает, какая доля ответов выдачи релевантна. **Выдача называется точной**, если все ответы релевантны.

Точность и полноту можно неформально называть «противоположными[1]» показателями качества. Легко построить алгоритм с максимальной полнотой: он все объекты относит к классу 1 (при запросе в поисковике выдаются ссылки на все ресурсы в интернете), но при этом точность может быть очень низкой. Нетрудно построить алгоритм со средней точностью близкой к 1: он относит к классу 1 только те объекты, в которых уверен, при этом полнота может быть низкая.

F_1-мера (F_1 score) является средним гармоническим точности и полноты, максимизация этого функционала приводит к одновременной максимизации этих двух «противоположных» показателей:

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2}{\frac{1}{\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})} + \frac{1}{\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})}} =$$

$$= \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$



Также рассматривают весовое среднее гармоническое точности (P) и полноты (R) – **F_β-меру (F_β score)**:

$$F_\beta = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}} = \frac{P \cdot R}{\alpha R + (1-\alpha) P} =$$

[1] Тут мы не нашли подходящего термина, максимизация одного показателя не обязательно приводит к минимизации другого, но чаще на практике именно так и происходит.

$$= \frac{1}{\alpha} \frac{P \cdot R}{R + \left(\frac{1}{\alpha} - 1\right) P}.$$

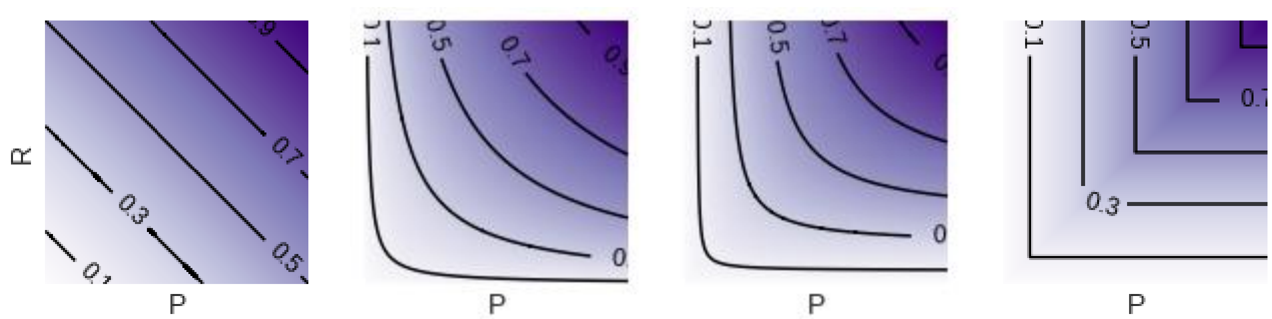
При этом β здесь не вес в среднем гармоническом, а

$$\beta^2 = \left(\frac{1}{\alpha} - 1\right),$$

таким образом:

$$F_\beta = (1 + \beta^2) \frac{P \cdot R}{R + \beta^2 P}.$$

Почему используется среднее гармоническое, понятно из рис. XY.4, на котором показаны линии уровня различных функций усреднения. Видно, что линии уровня среднего гармонического сильно похожи на «уголки», т.е. на линии функции $\min(P, R)$, что вынуждает при максимизации функционала сильнее «тянуть вверх» меньшее значение. Например, если точность очень мала, то увеличение полноты, пусть и в два раза, не сильно меняет значение функционала. Нагляднее это показано на рис. XY.5: $F_1 \leq 2/11$ при $P = 0.1$.



$(P+R)/2$ $\sqrt{P \cdot R}$ $2/(1/P + 1/R)$ $\min(P, R)$

Рис. XY.4. Линии уровня некоторых функций от двух переменных.

При использовании F_β-меры линии уровня «перекашиваются», один из критериев (точность или полнота) становится важнее при оптимизации, см. рис. XY.6.

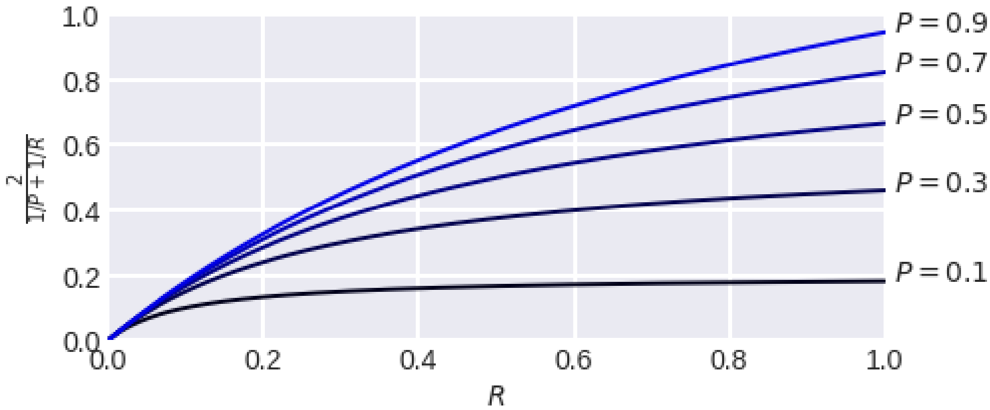
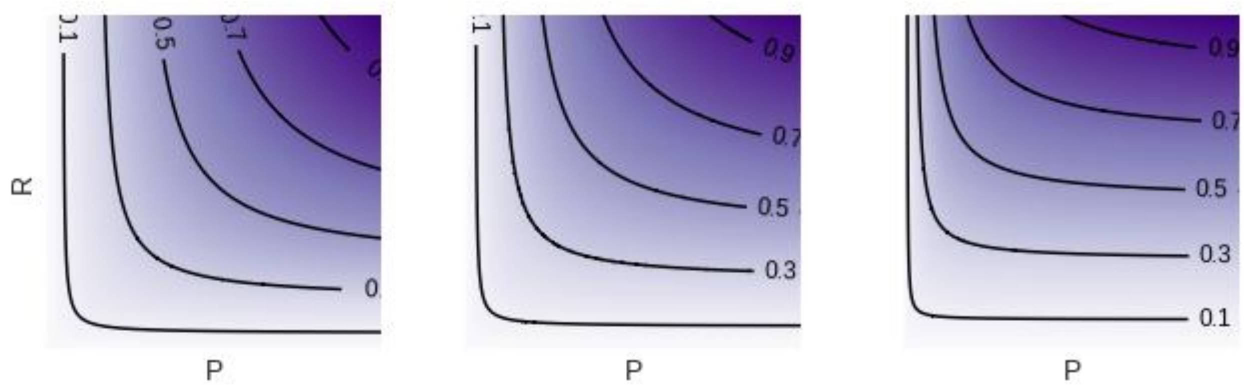


Рис. XY.5. Зависимость F-меры от полноты при фиксированной точности.



$1/(0.5/P + 0.5/R)$ $1/(0.3/P + 0.7/R)$ $1/(0.1/P + 0.9/R)$

Рис. XY.6. Линии уровня F-меры, $F_{1.53}$-меры и F_3-меры.

Из показателей качества, которые получаются из матрицы несоответствий, можно также отметить **специфичность (Specificity)** или **TNR – True Negative Rate**:

$$\mathrm{TNR} = \mathrm{Specificity} = R_0 = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$

	$a = 0$	$a = 1$
$y = 0$	TN	FP
$y = 1$		

это доля правильно классифицированных объектов негативного класса. Полноту иногда называют **чувствительностью (Sensitivity)** и используют в паре со специфичностью для оценки качества, также часто их усредняют[1] (об этом поговорим дальше). Оба функционала имеют смысл «процент правильно классифицируемых объектов одного из класса». Можно ввести понятие **полноты** R_k **для** k**-го класса**: это полнота, если считать класс k

[1] Усреднение чувствительности и специфичности часто применяют в медицинских задачах.

положительным, или, по-простому, доля объектов k-го, которую отнесли к этому классу, тогда

$$R_1 = R = \mathrm{Sensitivity},$$

$$R_0 = \mathrm{Specificity}.$$

Аналогично можно ввести **точность P_k для k-го класса**, например точность для класса 0 часто называют **обратной точностью (Inverse Precision, Negative Predictive Value)**:

$$\mathrm{NPV} = P_0 = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}.$$

	$a=0$	$a=1$
$y=0$	TN	
$y=1$	FN	

Также запомним **False Positive Rate (FPR, fall-out, false alarm rate):**

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{TNR}$$

– доля объектов негативного класса, которых мы ошибочно отнесли к положительному (этот показатель понадобится для понимания функционала AUC ROC, см. дальше).

Коэффициент Мэттьюса (MCC – Matthews correlation coefficient) равен

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \in [-1,+1],$$

его рекомендуют применять для несбалансированных выборок. Давайте разберёмся, что означает эта «сложная формула». Рассмотрим среднее геометрическое точности и полноты:

$$\sqrt{P\cdot R} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\cdot\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}} = \frac{\mathrm{TP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})}}.$$

Теперь возьмём среднее геометрическое точности и полноты класса 0 (т.е. считая это класс положительным), перемножив эти средние геометрические, получим значение на отрезке $[0,1]$, которое логично максимизировать:

$$\sqrt{P_1 R_1 P_0 R_0} = \frac{\mathrm{TP}\cdot\mathrm{TN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \in [0,1].$$

Получили одно из слагаемых в формуле MCC: произведение элементов на диагонали матрицы несоответствий делим на корень из произведения сумм столбцов и строк этой матрицы. Довольно очевидно[1], что если вычесть из этого произведение элементов антидиагонали, поделённое на такой же корень, то получим значение на отрезке на отрезке $[-1,+1]$ (оба крайних значения достижимы).

Как запомнить MCC: определитель матрицы несоответствий разделить на корень произведений сумм всех строк и столбцов.

Каппа Коэна (Cohen's Kappa) довольно часто используется в задачах классификации с дисбалансом классов. Идея довольно простая: поскольку использование точности (Accuracy) вызывает сомнение в задачах с сильном дисбалансом классов, надо её значения немного перенормировать. Делается это с помощью статистики **chance adjusted index**: мы точность нашего решения (Accuracy) пронормируем с помощью точности, которую можно было получить случайно ($\text{Accuracy}_{\text{chance}}$):

Стандартная идея: приведение значений на нужный отрезок.

$$\kappa = \frac{\text{Accuracy} - \text{Accuracy}_{\text{chance}}}{1 - \text{Accuracy}_{\text{chance}}}.$$

Под случайной здесь понимаем точность решения, которое получено из нашего случайной перестановкой ответов. Пусть у нас следующая матрица несоответствий:

	$a=0$	$a=1$
$y=0$	m_{00}	m_{01}
$y=1$	m_{10}	m_{11}

Формально наша точность:

$$\text{Accuracy} = \frac{m_{00} + m_{11}}{m}.$$

Теперь вычислим точность, которую можно было бы получить случайно:

$$\text{Accuracy}_{\text{chance}} = \frac{m_{00} + m_{01}}{m}\frac{m_{00} + m_{10}}{m} + \frac{m_{10} + m_{11}}{m}\frac{m_{01} + m_{11}}{m},$$

[1] Оставляем читателю как несложное упражнение.

здесь красным выделена вероятность (точнее её оценка) угадать класс 0, а синим – класс 1. Действительно, класс k угадывается, если алгоритм выдаёт метку k и объект действительно принадлежит этому классу. Предполагаем, что это независимые события (мы же хотим вычислить случайную точность). Вероятность принадлежности к классу k можно оценить по матрице несоответствий как долю объектов класса k. Аналогично, вероятность выдать метку оцениваем как долю таких меток в ответах построенного алгоритма.

Заметим, что $\kappa = 1$ при максимальной точности: $\mathrm{Accuracy} = 1$, а также $\kappa = 0$ при случайной точности: $\mathrm{Accuracy} = \mathrm{Accuracy}_{\mathrm{chance}}$. Можно надеяться, что для решения задачи $\kappa \in [0,1]$, т.к. решение должно быть лучше случайного.

Сбалансированная точность (Balanced Accuracy) является аналогом точности для задач с дисбалансом классов:

$$\mathrm{BA} = \frac{R_1 + R_0}{2} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right).$$

Для простоты запоминания, это среднее полноты всех классов (мы ещё вернёмся к этому определению), а для задачи с двумя классами – среднее чувствительности (Sensitivity) и специфичности (Specificity). Отметим, что чувствительность и специфичность можно, как полноту и точность, называть «противоположными» показателями. Легко сделать специфичность максимальной (равной 1), отнеся все объекты к классу 0, при этом будет нулевая чувствительность, и наоборот, если отнести все объекты к классу 1, то будет нулевая специфичность и единичная чувствительность.

Если в бинарной задаче классификации представителей двух классов в выборке примерно поровну, то

$$\mathrm{TP}+\mathrm{FN} \approx \mathrm{TN}+\mathrm{FP} \approx m/2$$

и сбалансированная точность примерно равна точности обычной (Accuracy):

$$\frac{1}{2}\left(\frac{\mathrm{TP}}{m/2} + \frac{\mathrm{TN}}{m/2}\right) = \frac{\mathrm{TP}+\mathrm{TN}}{m}.$$

Сравнение показателей качества для чёткой бинарной классификации

Рассмотрим модельную задачу, в которой плотности распределений классов на оценках, порождённых алгоритмом, линейные, см. рис. XY.7. Алгоритм выдаёт оценки принадлежности к классу 1 из отрезка $[0,1]$:

$$b(x_1),\ldots,b(x_m)\in[0,1],$$

именно на этом отрезке они линейные. Чёткая классификация получается после бинаризации оценок:

$$\forall i\in\{1,2,\ldots,m\}\quad a(x_i)=\begin{bmatrix}1, & b(x_i)\geq\theta,\\ 0, & b(x_i)<\theta,\end{bmatrix}$$

(если оценка выше фиксированного порога $\theta\in[0,1]$, то объект относим к классу 1, иначе – к классу 0).

На рис. XY.7 (в виде диаграммы рассеивания под графиком) показана конечная выборка, которая соответствует изображённым плотностям, мы же будем считать, что выборка «бесконечная[1]», поскольку плотности простые и позволяют в явном виде вычислить показатели качества. Будем считать, что классы равновероятны, т.е. наша бесконечная выборка сбалансирована[2].

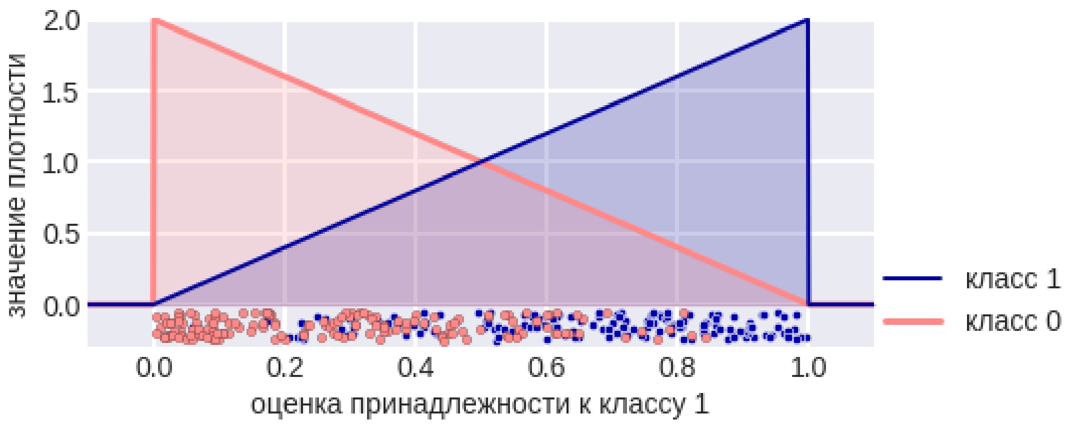


Рис. XY.7. Плотности распределения классов в модельной задаче классификации.

Как видно, алгоритм «растаскивает классы»: большинство объектов первого класса получила оценки близкие к 1, а большинство объектов нулевого класса – близкие к 0. Осталось выбрать оптимальный порог бинаризации $\theta\in[0,1]$.

[1] Строго говоря, дальше мы будем вычислять не показатели качества, а их матожидания.

[2] Выбранная задача очень удобна для исследования, будет использоваться и дальше.

Заметим, что подобные распределения возникают в задаче, показанной на рис. XY.8 (объекты лежат внутри квадрата $[0,1]\times[0,1]$, два класса разделяются диагональю квадрата), если алгоритм в качестве оценки выдаст значения первого признака (у нас мог быть доступен лишь один признак или алгоритм по каким-то причинам не использовал второй признак).

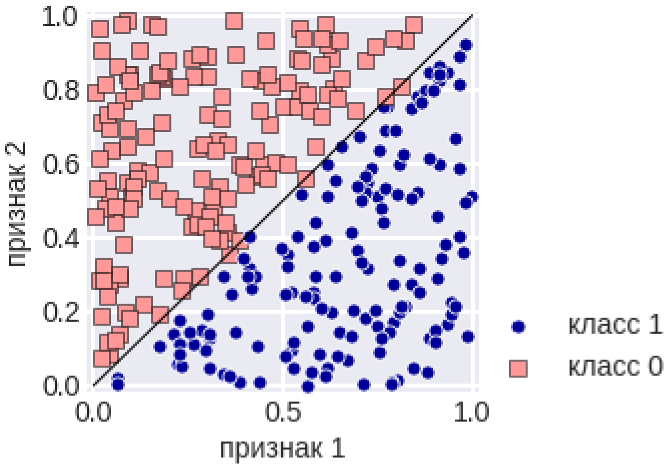


Рис. XY.8. Представление модельной задачи в признаковом пространстве.

Изобразив плотности немного по-другому, мы в явном виде можем вычислить элементы матрицы несоответствий при конкретном пороге бинаризации, см. рис. XY.9. Все они пропорциональны площадям выделенных зон (обратите внимание на масштаб осей): $\mathrm{FN}=\theta^2$ (это площадь треугольника с основанием равным θ и высотой 2θ) $\mathrm{TP}=1-\theta^2$, $\mathrm{FP}=(1-\theta)^2$, $\mathrm{TN}=1-(1-\theta)^2$.

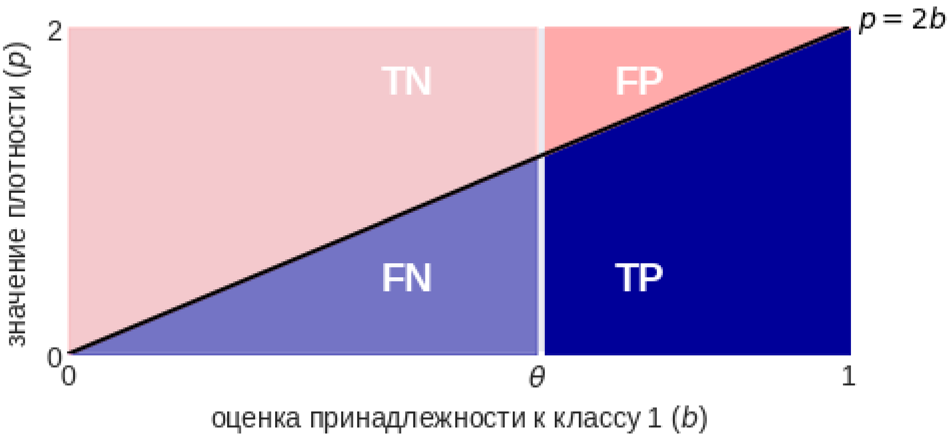


Рис. XY.9. Удобное изображение плотностей: пересекая рисунок прямой $a=\theta$, мы получаем значения плотностей в виде длин отрезков, которые попали на соответствующие цветные зоны: синюю – для класса 1, розовую – для класса 0.

Можно вывести[1] формулы для рассмотренных показателей качества в виде функций от порога бинаризации θ:

$$R = 1-\theta^2,\ P = (1+\theta)/2,\ F_1 = \frac{1-\theta^2}{1.5-\theta},$$

$$\text{BA=Accuracy} = \frac{(1-\theta^2)+(1-(1-\theta)^2)}{2} = \frac{1+2\theta-2\theta^2}{2},$$

$$\text{MCC} = \sqrt{\theta(1-\theta)},$$

$$\kappa = \frac{\dfrac{1+2\theta-2\theta^2}{2}-\dfrac{1}{2}}{1-\dfrac{1}{2}} = 2\theta(1-\theta).$$

Рассмотрим графики показателей качества как функций от порога бинаризации, см. рис XY.10. Заметим, что кроме F1-меры все они симметричны относительно вполне логичного для бинаризации порога $\theta = 0.5$.

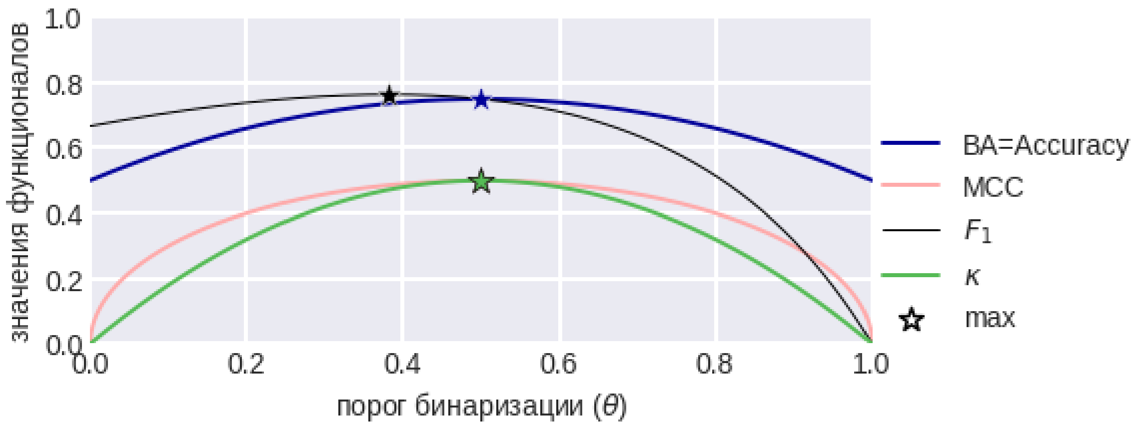


Рис. XY.10. Показатели качества как функции от порога бинаризации.

Для определения оптимального порога бинаризации при использовании F_1-меры, надо приравнять к нулю производную этого функционала

$$\frac{\partial F_1}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\frac{1-\theta^2}{1.5-\theta}\right) = 0,$$

получаем

[1] Попробуйте вывести эти формулы сами.

$$\theta_{\text{opt}} = \frac{3-\sqrt{5}}{2} \approx 0.38\,,$$

т.е. при использовании F_1-меры выгоднее больше объектов относить к классу 1 (чем при использовании остальных показателей качества).

Возникает естественный вопрос: на практике у нас нет бесконечных выборок[1], что изменится, если мы вычислим значения функционалов на конечной выборке, объекты которой сгенерированы в соответствии с указанными распределениями? Частично ответ на этот вопрос показан на рис. XY.11. Как видно, кривые довольно близки к теоретическим, при увеличении выборки в 10 раз практически совпадают.

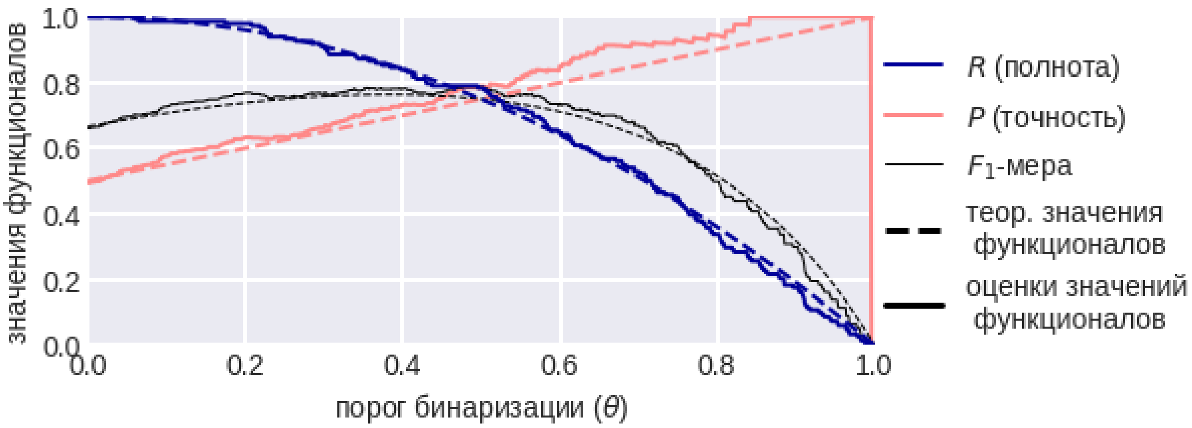


Рис. XY.11. Теоретические значения полноты, точности и F_1-меры, и значения, вычисленные по выборке из 300 объектов.

Теперь рассмотрим ситуацию неравновероятных классов, т.е. когда выборка несбалансированна[2]. На рис. XY.12 показаны графики функционалов в случае, когда класс 1 в два раза чаще встречается в выборке, чем класс 0. Обратите внимание, что все графики стали несимметричными, кроме графика сбалансированной точности – эта функция не зависит от пропорций классов! На рис. XY.13 также показаны оптимальные пороги для всех функционалов при всевозможных дисбалансах классов, а на рис. XY.14 их максимальные значения.

[1] Мы работаем не с матожиданиями показателей, а с показателями, вычисленными на конечных выборках.

[2] Здесь все аналитические выкладки опускаем.

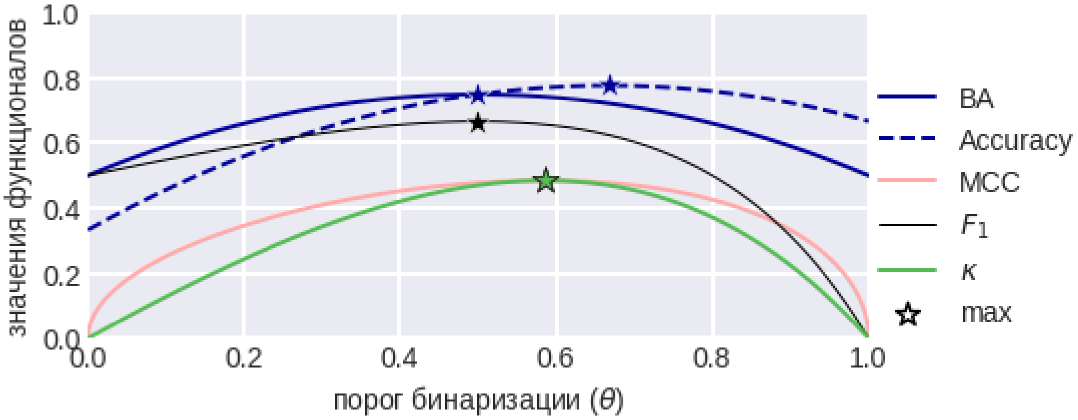


Рис. XY.12. Показатели качества как функции от порога бинаризации при дисбалансе классов (класс 1 в два раза вероятнее класса 0).

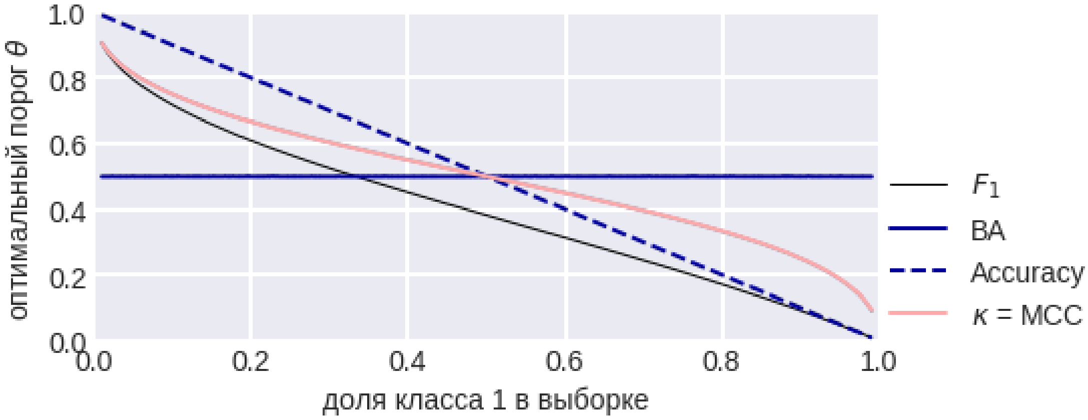


Рис. XY.13. Оптимальный порог бинаризации для разных показателей качества при изменении доли класса 1 в выборке.

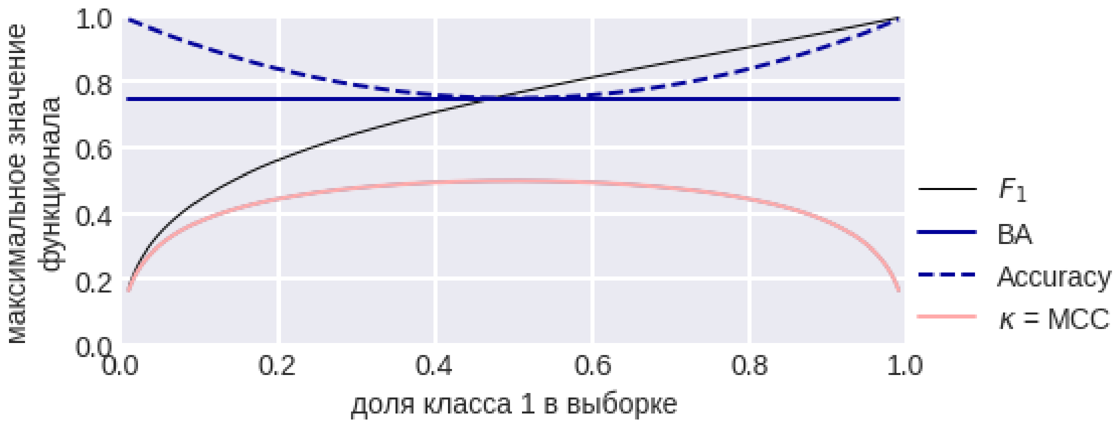


Рис. XY.14. Максимальные значения показателей качества при изменении доли класса 1 в выборке.

Из представленных графиков можно сделать много выводов. Например, заметим что на рис. XY.11 значения точности на конечной выборке больше отличаются от теоретических, особенно при больших порогах бинаризации. Это общее свойство точности (наблюдается не только в рассмотренной задаче) связано с тем, что при больших порогах бинаризации небольшая группа объектов относится алгоритмом к классу 1, а точность – доля объектов с меткой 1 среди них, поэтому оценка точности на таких маленьких выборках неустойчива. Например, если при небольшом понижении порога в группу попадает объект с меткой 0, то точность значительно падает. Остальной анализ графиков перенесён в теоретические вопросы.

Задачи и вопросы

- У какого показателя качества бинарной классификации самый маленький оптимальный порог бинаризации (в общем случае), почему?
- Какой показатель качества действительно имеет смысл использовать в задачах с сильным дисбалансом классов (заметим, что стандартные советы: BA, MCC, κ, F_1 обладают совершенно разными свойствами)?
- Какой «самый неустойчивый» из рассмотренных показателей (его значения на небольших выборках сильнее отличаются от вычисленных на достаточно больших)?
- Что изменится в примерах выше, если от линейных плотностей перейти к нормальным?
- Верно ли, что максимальное значение точности (т.е. значение точности при оптимальном выборе порога) всегда не меньше максимального значения сбалансированной точности?
- Выведите явные формулы для показателей качества в задаче с линейными плотностями и дисбалансом классов (придётся повозиться с формулами).
- Покажите, что показатель MCC может принимать все промежуточные значения отрезка $[0,1]$. Можно ли как-то интерпретировать то, что в числителе формулы для MCC стоит определитель матрицы несоответствий?
- Приведите пример задачи бинарной классификации, модели и способа настройки её параметров, при которых максимизация точности (Precision) не приводит к минимизации полноты (Recall).

Чёткая бинарная классификация: итоги

- Вся информация об ошибках содержится в матрице несоответствий: TP, TN, FP, FN, из указанных значений получаются различные показатели качества.

- Самый очевидный показатель качества, который плох при сильном дисбалансе классов, доля верных ответов:

$$\mathrm{Accuracy} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{FN} + \mathrm{TP} + \mathrm{FP}}.$$

- Полнота $R = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN})$ и точность $P = \mathrm{TP} / (\mathrm{TP} + \mathrm{FP})$ довольно естественны в информационном поиске, часто используют их среднее гармоническое – F_1-меру:

$$F_1 = \frac{2}{(1/P) + (1/R)}.$$

  Оптимизация F_1-меры приводит к тому, что алгоритм относит к классу 1 «достаточно много объектов». Есть весовое обобщение этого показателя – F_β-мера.

- В задачах с дисбалансом классов, кроме F_1-меры, используют также

$$\text{каппу Коэна: } \kappa = \frac{\mathrm{Accuracy} - \mathrm{Accuracy}_{\mathrm{chance}}}{1 - \mathrm{Accuracy}_{\mathrm{chance}}},$$

$$\text{коэффициент Мэттьюса: } \mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}},$$

$$\text{сбалансированную точность: } \mathrm{BA} = \frac{R_1 + R_0}{2}$$

  (первые два показателя схожи по своим свойствам, последний от них существенно отличается).

Спасибо за внимание к книге!
Замечания по содержанию, замеченные ошибки
и неточности можно написать в телеграм-чате
https://t.me/Dyakonovsbook